OXFORD
UNIVERSITY PRESS

Archives of Clinical Neuropsychology 36 (2021) 347–358

Archives of CLINICAL NEUROPSYCHOLOGY

Reliable change in cognition over 1 week in community-dwelling older adults: a validation and extension study

Dustin B. Hammers[1,2,*], Kayla R. Suhrie[1], Ava Dixon[1], Sariah Porter[1], Kevin Duff[1,2]

[1]*Department of Neurology, Center for Alzheimer's Care, Imaging, and Research, University of Utah*
[2]*Center on Aging, University of Utah*

*Corresponding author at: Department of Neurology, Center for Alzheimer's Care, Imaging and Research, University of Utah, 650 Komas Drive #106-A, Salt Lake City, UT 84108, USA. Tel.: 801-585-3929; Fax: 801-581-2483. *E-mail address*: dustin.hammers@hsc.utah.edu (Dustin B. Hammers).

Received 15 February 2019; revised 15 November 2019; Accepted 18 November 2019

Abstract

Objective: Reliable change methods can aid neuropsychologists in understanding if performance differences over time represent clinically meaningful change or reflect benefit from practice. The current study sought to externally validate the previously published standardized regression-based (SRB) prediction equations developed by Duff for commonly administered cognitive measures.

Method: This study applied Duff's SRB prediction equations to an independent sample of community-dwelling participants with amnestic mild cognitive impairment (MCI) assessed twice over a 1-week period. A comparison of MCI subgroups (e.g., single v. multi domain) on the amount of change observed over 1 week was also examined.

Results: Using pairwise *t*-tests, large and statistically significant improvements were observed on most measures across 1 week. However, the observed follow-up scores were consistently below expectation compared with predictions based on Duff's SRB algorithms. In individual analyses, a greater percentage of MCI participants showed smaller-than-expected practice effects based on normal distributions. In secondary analyses, smaller-than-expected practice effects were observed in participants with worse baseline memory impairment and a greater number of impaired cognitive domains, particularly for measures of executive functioning/speeded processing.

Conclusions: These findings help to further support the validity of Duff's 1-week SRB prediction equations in MCI samples and extend previous research by showing incrementally smaller-than-expected benefit from practice for increasingly impaired amnestic MCI subtypes.

Keywords: Reliable change; Assessment; Mild cognitive impairment

Introduction

A family of related statistical procedures, known as reliable change methods, has been developed to assist neuropsychologists in determining whether performance differences in serial assessments represent clinically meaningful change or reflect benefit from prior test exposure and practice effects (Hammers et al., 2015; Lezak et al., 2012). Such procedures consider the impact of a number of factors on retest performance, including differential practice effects and other systematic biases, regression to the mean, and measurement error (Chelune, 2003; Hinton-Bayre, 2010). The standardized regression-based (SRB) predicted difference method, first developed and characterized by McSweeny and colleagues (McSweeny et al., 1993), is a reliable change method that makes use of linear regression to predict retest scores (Time 2) for individuals based on their baseline (Time 1) performance. Specifically, a discrepancy change score (*z* score) is calculated from comparing an individual's predicted and observed Time 2 scores and dividing by the standard error of the estimate (SE_{Est}) of the regression model, to indicate the degree

doi:10.1093/arclin/acz076 Advance Access publication 6 February 2020

of deviation from expectation. This approach is capable of incorporating not only baseline performance into the predicted Time 2 score but also *personalized* practice effects and other potential and unique sources of systematic bias (e.g., regression to the mean, patient characteristics). Given the utility of serial assessment with older adults, the number of SRB predicted difference equations in the literature for common neuropsychological tests continues to grow (Attix et al., 2009; Crockford et al., 2018; Duff et al., 2004; Duff et al., 2005; Duff et al., 2010a; Gavett et al., 2015; Rinehardt et al., 2010; Sanchez-Benavides et al., 2016; Stein et al., 2010).

Traditionally, when examining change over longer periods of time, discrepancy change scores—or z scores—below -1.645 represent "decline," whereas z scores > 1.645 reflect "improvement" and z scores between $+/-1.645$ indicate stability. This z score cut-off of 1.645 is used because it equates to an α value of $\alpha = .10$, indicating a 90% confidence interval of stability (McSweeny et al., 1993). If the z scores were normally distributed, then one would expect that 5% of participants would show "decline," 90% would remain "stable," and 5% would "improve" beyond expectation. However, when examining change over shorter periods of time, follow-up scores may be better than baseline scores, yet still worse than predictions based on the SRB algorithms. In this case, the use of the term "decline" might be replaced with "smaller-than-expected improvement" when z scores are < -1.645. Conversely, an individual could show "larger-than-expected improvement" ($z > 1.645$) or "typical improvement" (z scores between $+/-$ 1.645). Further, given that our sample population is not expected to display acute changes in cognition/treatment response over such a short duration, it is proposed that these findings reflect "smaller-than-expected" or "larger-than-expected" *practice effects*.

Duff (2014) created regression-based prediction equations for several cognitive tests that are commonly administered both clinically and in research settings, including the Hopkins Verbal Learning Test—Revised (HVLT-R; Brandt & Benedict, 2001), the Brief Visual Memory Test—Revised (BVMT-R; Benedict, 1997), Symbol Digit Modality Test (SDMT; Smith, 1973), and the Trail Making Test Parts A and B (TMT-A and TMT-B; Reitan, 1992). These prediction equations were developed on 167 community-dwelling older adults, of whom 93 were classified as cognitively intact and 74 were classified as having mild cognitive impairment (MCI); both cognitively intact and MCI samples were included in the study "to increase the range of test scores" (Duff, 2014, p. 716). This data was collected over two test administrations 1 week apart, which permits examination of the impact of short-term practice effects on cognitive performance.

Unfortunately, there have been limited attempts to validate these SRB prediction equations. In Duff's original 2014 publication, there was no internal validation of the change formulae. Subsequently, Duff and colleagues (Duff et al., 2018; Duff et al., 2017) applied these prediction equations to relatively small samples (n = 25–58) of community-dwelling older adults with varying levels of cognitive abilities who were tested twice across 1 week. Both studies found that the more impaired participants tended to have smaller-than-expected practice effects, though their sample sizes prevented findings from being extrapolated with confidence (Faber & Fonseca, 2014). Therefore, additional research is needed to thoroughly validate these SRB prediction equations in larger sample sizes. Consequently, the aims of the current study were to both replicate the observed validity of these SRB prediction equations using a larger sample of community-dwelling older adults with MCI and also extend the examination of criterion validity by comparing subgroups of the sample (single-domain amnestic MCI [aMCI-S] v. multi-domain amnestic MCI [aMCI-M]) on the amount of change observed over 1 week. Based on Duff's previous validation sample (Duff et al., 2017), it was hypothesized that the application of these prediction equations to an independent sample of older adults with MCI would result in a greater proportion of participants failing to benefit from practice on these cognitive measures over 1 week than expected. As research has suggested that MCI patients who fail to benefit from practice have worse outcomes (Duff et al., 2011; Hassenstab et al., 2015; Machulda et al., 2013), worse response to intervention (Duff et al., 2010b), and greater risk of Alzheimer's-related pathology (Duff et al., 2018; Duff et al., 2014; Galvin et al., 2005; Mormino et al., 2014) than those who show improvements on retesting, further validation of these SRB prediction equations could possess diagnostic and prognostic value and inform treatment recommendations in patients with MCI.

Method

Participants

One-hundred forty-three community-dwelling older adults were recruited from either a cognitive disorders clinic (61%) or senior centers and independent living facilities (39%). Their mean age was 75.5 (standard deviation [SD] = 6.1, range = 65–91) years old, and they averaged 16.2 (SD = 2.9, range = 12–20) years of education. The sample of participants was evenly divided by sex (50.3% female), and the majority were Caucasian (97.9%). Premorbid intellect at baseline was average according to the Wide Range Achievement Test—fourth edition (WRAT-4; Wilkinson & Robertson, 2006) Reading subtest (standard score: M = 108.2, SD = 8.8, range = 85–145). For inclusion in the study, all participants from this sample were classified as having either aMCI-S or aMCI-M. Classification of participants from this sample has been described previously (Duff et al.,

Table 1. Demographic characteristics of the current validation sample

Variables	Mean (*SD*)	Range
n	143	
Age (years)	75.5 (6.1)	65–91
Education (years)	16.2 (2.9)	12–20
Gender (*n*)		
Males	71	
Females	72	
Race (*n*)		
African American	1	
Hispanic/Latino American	1	
Native American	1	
White, Non-Hispanic	140	
Test interval (days)	7.2 (0.9)	6–13
WRAT-4 premorbid intellect (SS)	108.2 (8.8)	85–145
RBANS indexes (SS)		
Immediate memory	81.9 (16.7)	44–114
Visuospatial/constructional	97.7 (15.5)	62–136
Language	90.9 (12.3)	40–122
Attention	96.1 (15.2)	64–132
Delayed memory	77.7 (21.0)	40–122
Total scale	85.1 (13.2)	50–121

Notes: *SD* = standard deviation, WRAT-4 Premorbid Intellect = Wide Range Achievement Test—fourth edition Reading Subtest, RBANS = Repeatable Battery for the Assessment of Neuropsychological Status, SS = Standard Score.

2017). Briefly, participants were classified as amnestic MCI by participant and knowledgeable informant report and a baseline cognitive evaluation described below. Cognitive impairment for a domain was defined as a significant discrepancy (e.g., 1.5 *SD*) between cognitive performance and an estimate of premorbid intellect. As can be observed in Table 1, on average, the sample displayed below expectation abilities for baseline immediate and delayed memory skills, particularly after considering their strong premorbid intellect, though their cognition was otherwise generally intact. General inclusion criteria for the study involved being aged 65 years or older and functionally independent (according to participant and/or knowledgeable informant), along with possessing adequate vision, hearing, and motor abilities to complete the cognitive evaluation. General exclusion criteria included neurological conditions likely to affect cognition, dementia, major psychiatric condition, current severe depression, substance abuse, anti-convulsant or anti-psychotic medications, or residence in a skilled nursing or living facility.

Procedure

All procedures were approved by the local Institutional Review Board before the study commenced. All participants provided informed consent before completing any procedures. The following measures were administered at a baseline visit:

- HVLT-R (Brandt & Benedict, 2001) is a verbal memory task with 12 words learned over three trials, with the correct words summed for the Total Recall score (range = 0–36). The Delayed Recall score is the number of correct words recalled after a 20–25-minute delay (range = 0–12). For all HVLT-R scores, higher values indicate better performance.
- BVMT-R (Benedict, 1997) is a visual memory task with six geometric designs in six locations on a card learned over three trials, with correct designs and locations summed for the Total Recall score (range = 0–36). The Delayed Recall score is the number of correct designs and locations recalled after a 20–25-minute delay (range = 0–12). For all BVMT-R scores, higher values indicate better performance.
- SDMT (Smith, 1973) is a divided attention and psychomotor speed task, with the number of correct responses in 90 seconds being the total score (range = 0–110), and higher values indicate better performance.
- TMT-A and TMT-B (Reitan, 1992) are tests of visual scanning/processing speed and set shifting/complex mental flexibility, respectively. For each part, the score is the time to complete the task (range = 0–180 s for TMT-A, and range = 0–300 s for TMT-B), and higher values indicate poorer performance.
- WRAT-4 Reading (Wilkinson & Robertson, 2006) is used as an estimate of premorbid intellect, in which an individual attempts to pronounce irregular words. The score is normalized to Standard Scores (M = 100, SD = 15) relative to age-matched peers, and higher values indicate better performance.
- Repeatable Battery for the Assessment of Cognition (RBANS; Randolph, 2012) is a neuropsychological test battery comprising 12 subtests that are used to calculate Index scores for domains of immediate memory, visuospatial/constructional,

attention, language, delayed memory, and global neuropsychological functioning. The index scores utilize age-corrected normative comparisons from the test manual to generate standard scores ($M = 100$, $SD = 15$), with higher scores indicating better cognition.

After approximately 1 week ($M = 7.2$ days, $SD = 0.9$, range = 6–13), the HVLT-R, BVMT-R, SDMT, TMT-A, and TMT-B were repeated. The same form of each test was used to maximize practice effects for these repeated cognitive tasks. The RBANS and WRAT-4 were only administered at baseline, and baseline scores from all tests were used in the classification of MCI. With the exception TMT-B possessing three missing values, no other variables of interest possessed missing data.

Analyses

Pairwise baseline versus 1-week analyses. To approximate a traditional evaluation of change over time (comparison of Time 1 and Time 2 scores) without controlling for practice effects or participant variables, pair-wise *t* tests were conducted to compare observed baseline and observed 1-week scores for each of the repeated measures in the cognitive battery (HVLT-R, BVMT-R, SDMT, TMT-A, TMT-B).

SRB group analyses. Previously published SRB prediction equations for each of the measures in the cognitive battery were applied to the current sample's baseline and 1-week scores. As has been described previously (Duff, 2014), the SRB prediction algorithms were calculated from a developmental sample using stepwise multiple-regression analyses to maximize the prediction of performance for each repeated measure in the cognitive battery. Specifically, the combination of demographic variables (e.g., age, education, gender), test interval, and baseline test score was used to predict the respective test score at follow-up 1 week later.

Following the application of these SRB prediction equations to the current MCI sample, a *z* score was calculated for each participant, which reflects a normalized deviation of change for an individual participant. Specifically, the observed 1-week score (T_2) was compared with the predicted 1-week score (T_2'), normalized by the SE_{est} (i.e., $z = (T_2 - T_2')/SE_{est}$). *z* scores for each repeated measure in the cognitive battery were then compared with expectation ($z = 0$) based on the normal distribution of *z* scores using a one-sample *t* test.

Individual distribution analyses. Further, the resultant *z* scores were trichotomized into "smaller-than-expected practice effects" (*z* score < -1.645), "expected practice effects" (*z* score falling between $+/-1.645$), or "greater-than-expected practice effects" (*z* score > 1.645) for all measures in the repeated battery except for TMT-A and TMT-B, which used reversed scoring. As indicated previously, if the *z* scores were normally distributed, then one would expect that 5% of participants would show "smaller-than-expected practice effects," 90% would indicate "expected practice effects," and 5% would reflect "greater-than-expected practice effects." Using this trichotomization, individual chi-square analyses were conducted for each measure in the repeated cognitive battery to determine if the observed distribution of participants deviated significantly from the expected distribution based on the normal distribution of *z* scores.

Secondary analyses. Finally, the participants in this sample were further sub-categorized into MCI amnestic subtypes based on their baseline performances on non-memory cognitive domains (attention, visuospatial, language, and executive functioning). As described above, cognitive impairment for a domain was defined as a significant discrepancy (e.g., 1.5 *SD*) between cognitive performance on the baseline cognitive measures and an estimate of premorbid intellect. Participants impaired only on memory measures were classified as aMCI-S, whereas participants impaired on both memory and non-memory domains were classified as aMCI-M. Participants were subsequently grouped based on the number of non-memory domains in which they were impaired, out of a possible four domains in total. Due to the distribution of the samples, the three participant groups included aMCI-S (memory impairment only), aMCI-M1/2 (memory impairment plus impairment on 1–2 non-memory domains), and aMCI-M3/4 (memory impairment plus impairment on 3–4 non-memory domains). Following this classification, *z* scores were compared for each repeated measure in the cognitive battery between the aMCI-S and various aMCI-M groups using analysis of variance (ANOVA), with least squared difference (LSD) post-hoc analyses performed to examine individual subgroup comparisons. Further, individual chi-square analyses were conducted for dichotomized *z* scores to determine if differences were observed between the aMCI-S and various aMCI-M groups in the distribution of participants who displayed smaller-than-expected practice effects versus expected/greater-than-expected practice effects after 1 week. It is important to note that a high number of cells in this latter analysis contained no participants who possessed greater-than-expected practice effects, consequently the most parsimonious solution was to modify the methods from the primary analyses and combine those who displayed expected

Table 2. Baseline, observed and predicted 1-week cognitive scores, standardized z scores, and p values for difference from expectation ($z = 0$) based on the normal distribution of z scores in MCI participants

Measures	Observed baseline	Observed 1-week	Predicted 1-week	z score	p value
Hopkins Verbal Learning Test—Revised					
Total recall	18.0 (5.1)	21.6 (6.5)	23.4 (4.0)	−0.48 (1.1)	.001
Delayed recall	3.5 (3.3)	5.8 (3.7)	7.6 (1.7)	−0.94 (1.3)	.001
Brief Visual Memory Test—Revised					
Total recall	10.5 (5.7)	17.4 (8.5)	20.2 (5.8)	−0.53 (1.0)	.001
Delayed recall	3.7 (2.9)	6.2 (3.5)	6.4 (2.1)	−0.07 (1.2)	.48
Symbol Digit Modality Test	35.3 (9.7)	35.8 (10.6)	39.0 (8.8)	−0.65 (1.1)	.001
Trail Making Test					
Part A	43.8 (15.8)	41.2 (15.1)	38.3 (11.7)	0.28 (1.1)	.002
Part B	144.3 (79.5)	129.9 (77.9)	114.5 (48.1)	0.46 (1.5)	.001

Note: MCI = mild cognitive impairment, p value = significance of one-sample t tests examining whether z scores differed from expectation ($z = 0$) based on the normal distribution of z scores.

practice effects and greater-than-expected practice effects, resulting in dichotomized (instead of trichotomized) z scores. Finally, quartiles of performance were calculated for both delayed memory measures (HVLT-R Delayed Recall and BVMT-R Delayed Recall) at baseline across all participants in the sample (regardless of MCI subgroupings). Specifically, the first quartile group reflected an HVLT Delayed raw score of 0 ($n = 43$) or a BVMT Delayed raw score of 0–3 ($n = 39$), the second quartile group reflected an HVLT Delayed raw score of 1–3 ($n = 38$) or a BVMT Delayed raw score of 4–6 ($n = 36$), the third quartile group reflected an HVLT Delayed raw score of 4–6 ($n = 29$) or a BVMT Delayed raw score of 7–9 ($n = 36$), and the fourth quartile group reflected an HVLT Delayed raw score of 7–12 ($n = 23$) or a BVMT Delayed raw score of 10–12 ($n = 32$). Following this categorization, z scores were compared for each delayed memory measure as a function of baseline memory performance using ANOVA, with LSD post-hoc analyses to examine individual quartile comparisons.

Measures of effect size were expressed throughout as Cohen's d values (t test analyses) and eta squared (η^2) values (ANOVA analyses) for continuous data, and *Phi* coefficients for categorical data. A two-tailed alpha level was set at .05 for all statistical analyses.

Results

Pairwise Baseline Versus 1-Week Analyses

When examining change over time using a traditional method of comparing observed baseline and observed 1-week scores for each of the repeated measures in the cognitive battery (HVLT-R, BVMT-R, SDMT, TMT-A, TMT-B; see Table 2 for means) in this MCI sample, significant differences were observed for the HVLT-R Total Recall, $t(142) = -11.25$, $p = .001$, $d = -1.92$, HVLT-R Delayed Recall, $t(142) = -13.33$, $p = .001$, $d = -2.24$, BVMT-R Total Recall, $t(142) = -16.50$, $p = .001$, $d = -2.77$, BVMT-R Delayed Recall, $t(142) = -15.01$, $p = .001$, $d = -2.52$, TMT-A, $t(142) = 2.71$, $p = .008$, $d = 0.45$, and TMT-B, $t(139) = 3.31$, $p = .001$, $d = 0.56$. Specifically, scores were better at observed 1-week than at observed baseline for all six measures. No difference was observed for the SDMT, $t(142) = -1.23$, $p = .22$, $d = -0.21$.

SRB Group Analyses

SRB prediction equations for each of the repeated measures in the cognitive battery from Duff (2014) were applied to the current sample of MCI participants. As seen in Table 2, when comparing z scores for each repeated measure to expectation ($z = 0$) based on the normal distribution of z scores using one-sample t tests, significant differences were observed for six of the seven measures administered twice over 1 week. As a reminder, when z scores were significantly larger than zero for all measures but the TMT subtests, the current validation sample exceeded expectations based on Duff's developmental sample and reflected greater-than-expected practice effects over 1 week. Conversely, when z scores were significantly smaller than zero for these same measures (excluding TMT subtests), the current validation sample fell below expectations based on Duff's developmental sample and subsequently reflected smaller-than-expected practice effects over 1 week. Because higher values reflect worse performance for TMT-A and TMT-B, for those tests significantly smaller-than-zero z scores reflected exceeding expectations, whereas significantly larger-than-zero z scores represented falling below expectations.

Examining the repeated measures in the cognitive battery more specifically, this MCI sample displayed lower z scores than expected on subtests of HVLT-R Total Recall, $t(142) = -5.33$, $p = .001$, $d = -0.89$, HVLT-R Delayed Recall, $t(142) = -8.42$,

Table 3. Percentage of MCI sample that displayed smaller-than-expected practice effects, expected practice effects, or greater-than-expected practice effects based on standardized regression-based methodology

Measures	Practice effect			*p* value
	Smaller-than- expected	Expected	Greater-than-expected	
Hopkins Verbal Learning Test—Revised				
Total recall	14	85	1	.001
Delayed recall	29	71	0	.001
Brief Visual Memory Test—Revised				
Total recall	12	84	4	.006
Delayed recall	10	83	7	.04
Symbol Digit Modality Test	16	81	3	.001
Trail Making Test				
Part A	10	86	4	.07
Part B	17	79	4	.001

Note: MCI = mild cognitive impairment, *p* value = significance of chi-square tests between observed distribution and expected distribution based on the normal curve distribution of *z* scores (5% display smaller-than-expected practice effects, 90% display expected practice effects, 5% display greater-than-expected practice effects).

$p = .001$, $d = -1.41$, BVMT-R Total Recall, $t(142) = -6.29$, $p = .001$, $d = -1.05$, and SDMT, $t(142) = -7.18$, $p = .001$, $d = -1.21$, and significantly higher *z* scores than expected for TMT-A, $t(142) = 3.16$, $p = .002$, $d = 0.53$, and TMT-B, $t(139) = 3.68$, $p = .001$, $d = 0.62$. No difference was observed for the BVMT-R Delayed Recall *z* score relative to expectation based on the normal curve, $t(142) = -0.71$, $p = .48$, $d = -0.12$.

Individual Distribution Analyses

When examining the distribution of individual MCI participants that displayed "smaller-than-expected practice effects" (*z* score < -1.645 for HVLT-R, BVMT-R, and SDMT; *z* score > 1.645 for TMT-A and TMT-B), "expected practice effects" (*z* score falling between $+/-1.645$), or "greater-than-expected practice effects" (*z* score > 1.645 for HVLT-R, BVMT-R, and SDMT; *z* score < -1.645 for TMT-A and TMT-B) between baseline and 1-week administrations of the repeated cognitive battery, the majority of participants exhibited the expected level of improvement or practice effect (81.3% of participants; see Table 3). However, greater proportions of individuals displayed smaller-than-expected practice effects over 1 week than expected based on normal distributions for most measures: HVLT-R Total Recall, χ^2 (2) = 19.68, $p = .001$, *Phi* = .37 (14% of participants), HVLT-R Delayed Recall, χ^2 (2) = 124.21, $p = .001$, *Phi* = .93 (29% of participants), BVMT-R Total Recall, χ^2 (2) = 10.40, $p = .006$, *Phi* = .27 (12% of participants), BVMT-R Delayed Recall, χ^2 (2) = 6.34, $p = .04$, *Phi* = .21 (10% of participants), SDMT, χ^2 (2) = 25.90, $p = .001$, *Phi* = .43 (16% of participants), and TMT-B, χ^2 (2) = 30.34, $p = .001$, *Phi* = .47 (17% of participants). Although a trend was observed, no significant difference in performance distribution was seen relative to expectation for TMT-A, χ^2 (2) = 5.38, $p = .07$, *Phi* = .19. On no measure in the cognitive battery did greater proportions of individuals with MCI possess greater-than-expected practice effects over 1 week than anticipated based on the normal distribution of *z* scores (greater-than-expected practice effects was generally around the expected 5% value for each measure).

Secondary Analyses

As indicated above, participants were further sub-categorized into MCI amnestic subtypes based on their baseline performances on non-memory cognitive domains. Level of impairment was grouped into aMCI-S (memory impairment only; $n = 24$), aMCI-M1/2 (memory impairment plus impairment on 1–2 non-memory domains; $n = 73$), and aMCI-M3/4 (memory impairment plus impairment on 3–4 non-memory domains; $n = 46$). There were no differences in age, $F(2, 140) = 0.89$, $p = .42$, $\eta^2 = .012$, education, $F(2, 140) = 0.31$, $p = .74$, $\eta^2 = .004$, or gender, $\chi^2(2) = 1.00$, $p = .61$, *Phi* = .08, for the three groups.

Following this classification, *z* scores, which are a metric of the discrepancy between observed and predicted 1-week scores, were compared for each measure in the cognitive battery between the aMCI-S and various aMCI-M groups. Results indicated that significant differences were observed among the three MCI groups on BVMT-R Total Recall, $F(2, 140) = 3.71$, $p = .03$, $\eta^2 = .050$, and BVMT-R Delayed Recall, $F(2, 140) = 3.24$, $p = .04$, $\eta^2 = .044$. As can be observed in Table 4, post-hoc analyses revealed that lower *z* scores (worse observed performance compared to prediction) were observed for the more impaired groups (aMCI-M3/4 < both aMCI-S and aMCI-M1/2, $p = .03$ and $p = .02$, respectively, for BVMT-R Total Recall; aMCI-M3/4 < both aMCI-S and aMCI-M1/2, $p = .04$ and $p = .03$, respectively, for BVMT-R Delayed Recall). Additionally, non-significant trends

Table 4. Standardized z score values for each of the MCI impairment groups

Measures	aMCI-S ($n = 24$)	aMCI-M1/2 ($n = 73$)	aMCI-M3/4 ($n = 46$)	p value
Hopkins Verbal Learning Test—Revised				
Total recall	−0.19 (1.0)	−0.40 (1.0)	−0.76 (1.1)	.07
Delayed recall	−0.68 (1.3)	−0.83 (1.4)	−1.24 (1.2)	.15
Brief Visual Memory Test—Revised				
Total recall	−0.29 (0.9)	−0.40 (1.0)	−0.85 (1.0)	.03
Delayed recall	0.19 (1.0)	0.07 (1.2)	−0.43 (1.2)	.04
Symbol Digit Modality Test	−0.53 (0.9)	−0.51 (1.0)	−0.94 (1.3)	.09
Trail Making Test				
Part A	−0.02 (0.9)	0.22 (1.0)	0.55 (1.2)	.08
Part B	0.27 (1.2)	0.32 (1.3)	0.81 (1.9)	.17

Note: MCI = mild cognitive impairment, z score = observed minus predicted 1-week scores/standard error of the estimate of the regression. Lower z scores reflect worse performance for all measures except Trail Making Test, where high z scores reflect worse performance, aMCI-S = amnestic MCI single-domain, aMCI-M1/2 = amnestic MCI multi-domain including 1–2 non-memory impaired domains, aMCI-M3/4 = amnestic MCI multi-domain including 3–4 non-memory impaired domains, p value = significance of analysis of variances examining the difference in z scores between the three MCI impairment groups.

Table 5. Percentage of MCI sub-categorization samples that displayed smaller-than-expected practice effects, expected practice effects, or greater-than-expected practice effects based on standardized regression-based methodology

Measures	aMCI-S ($n = 24$)		aMCI-M1/2 ($n = 73$)		aMCI-M3/4 ($n = 46$)	
	Practice effect		Practice effect		Practice effect	
	Smaller-than-expected	Expected/greater-than-expected	Smaller-than-expected	Expected/greater-than-expected	Smaller-than-expected	Expected/greater-than-expected
Hopkins Verbal Learning Test—Revised						
Total recall	4	96	12	88	22	78
Delayed recall	21	79	27	73	37	63
Brief Visual Memory Test—Revised						
Total recall	8	92	8	92	20	80
Delayed recall	4	96	7	93	17	83
Symbol Digit Modality Test**	12	88	10	90	28	72
Trail Making Test						
Part A	0	100	10	90	15	85
Part B**	13	87	11	89	30	70

Note: MCI = mild cognitive impairment, aMCI-S = amnestic MCI single-domain, aMCI-M1/2 = amnestic MCI multi-domain including 1–2 non-memory impaired domains, aMCI-M3/4 = amnestic MCI multi-domain including 3–4 non-memory impaired domains.

** $p < .05$.

were observed for HVLT-R Total Recall, $F(2, 140) = 2.65$, $p = .07$, $\eta^2 = .036$, HVLT-R Delayed Recall, $F(2, 140) = 1.94$, $p = .15$, $\eta^2 = .027$, SDMT, $F(2, 140) = 2.50$, $p = .09$, $\eta^2 = .035$, TMT-A, $F(2, 140) = 2.54$, $p = .08$, $\eta^2 = .035$, and TMT-B, $F(2, 137) = 1.78$, $p = .17$, $\eta^2 = .025$, with similar observations that the more impaired groups possessed worse z scores.

Further, individual chi-square analyses were conducted for dichotomized z scores to determine if differences were observed between the aMCI-S and various aMCI-M groups in the distribution of participants who displayed smaller-than-expected practice effects versus expected/greater-than-expected practice effects after 1 week (see Table 5). Overall, greater proportions of individuals exhibited smaller-than-expected practice effects over 1 week for the aMCI-3/4 group than the aMCI-S group for SDMT, $\chi^2(2) = 7.56$, $p = .02$, $Phi = .23$, and TMT-B, $\chi^2(2) = 7.00$, $p = .03$, $Phi = .22$. For both tasks, 12–13% of aMCI-S participants displayed smaller-than-expected practice effects relative to 28–30% of aMCI-M3/4 participants. Similar non-significant trends were observed for HVLT-R Total Recall, $\chi^2(2) = 4.39$, $p = .11$, $Phi = .18$, BVMT-R Total Recall, $\chi^2(2) = 3.82$, $p = .15$, $Phi = .16$, BVMT-R Delayed Recall, $\chi^2(2) = 4.58$, $p = .10$, $Phi = .18$, and TMT-A, $\chi^2(2) = 4.14$, $p = .13$, $Phi = .17$. No differences in distributions were observed for HVLT-R Delayed Recall between impairment groups, $\chi^2(2) = 2.26$, $p = .32$, $Phi = .13$, though this was likely due to high proportions of participants performing below expectation for this measure across all three groups.

Finally, z scores for HVLT-R Delayed Recall and BVMT-R Delayed Recall were compared across MCI subgroupings as a function of performance ability for each task at baseline. Results indicated that significant differences were observed among the four quartile groups on HVLT-R Delayed Recall, $F(3, 139) = 29.29$, $p < .001$, $\eta^2 = .387$, and BVMT-R Delayed Recall,

Table 6. Standardized *z* score values as a function of baseline performance for each respective memory measure across the current MCI sample

Measures	First quartile	Second quartile	Third quartile	Fourth quartile	*p* value
HVLT-R delayed recall	−2.14 (1.1)	−0.70 (1.2)	−0.52 (1.1)	−0.01 (0.7)	<.001
BVMT-R delayed recall	−1.39 (0.6)	−0.26 (0.7)	0.48 (0.9)	1.14 (0.7)	<.001

Note: *z* score = observed minus predicted 1-week scores/standard error of the estimate of the regression, MCI = mild cognitive impairment, HVLT-R = Hopkins Verbal Learning Test—Revised, BVMT-R = Brief Visual Memory Test—Revised, *p* value = significance of analysis of variances examining the difference in z scores between the four quartile groups. First quartile reflects HVLT Delayed raw score of 0 ($n = 43$) or BVMT Delayed raw score of 0–3 ($n = 39$), second quartile reflects HVLT Delayed raw score of 1–3 ($n = 38$) or BVMT Delayed raw score of 4–6 ($n = 36$), third quartile reflects HVLT Delayed raw score of 4–6 ($n = 29$) or BVMT Delayed raw score of 7–9 ($n = 36$), and fourth quartile reflects HVLT Delayed raw score of 7–12 ($n = 23$) or BVMT Delayed raw score of 10–12 ($n = 32$).

$F(3, 139) = 79.77$, $p < .001$, $\eta^2 = .633$. As can be observed in Table 6, post-hoc analyses revealed that lower *z* scores (worse observed performance compared to prediction) were generally observed for the more impaired groups (first quartile < second quartile < third quartile < fourth quartile for BVMT-R Delayed Recall, *p*s < .001; first quartile < all other groups and second quartile < fourth quartile for HVLT-R Delayed Recall, *ps* < .001).

Discussion

The current study sought to examine the validity of previously published SRB predicted difference equations (Duff, 2014) for a set of commonly administered cognitive measures, including the HVLT-R, BVMT-R, SDMT, TMT-A, and TMT-B using independent samples of amnestic MCI community-dwelling older adults assessed twice over a 1-week period. While Duff and colleagues (Duff et al., 2018; Duff et al., 2017) have attempted to externally validate these SRB prediction equations previously, their relatively small samples of participants with MCI limited the generalizability of those findings. In addition, the current study extended previous research by comparing subgroups of the sample (aMCI-S v. aMCI-M) on the amount of change observed over 1 week, which to our knowledge is the first study to do so.

For our current validation sample of MCI participants, when comparing observed test scores at baseline and 1-week, large and statistically significant improvements in performance were observed across most measures administered (HVLT-R Total Recall, HVLT-R Delayed Recall, BVMT-R Total Recall, BVMT-R Delayed Recall, TMT-A, and TMT-B; Cohen's *d*s = | 0.45–2.77 |). In contrast, when applying Duff's (2014) SRB prediction equations to baseline performance on these measures, the observed 1-week scores for our sample of MCI participants were consistently *below expectation* compared with predictions (HVLT-R Total Recall, HVLT-R Delayed Recall, BVMT-R Total Recall, SDMT, TMT-A, and TMT-B; Cohen's *d*s = | 0.53–1.41 |). Additionally, when examining the distributions of participants that displayed smaller-than-expected, expected, or greater-than-expected practice effects in our sample, greater proportions of individuals performed worse than expected based on normal distributions for several measures (14 and 29% of participants displayed smaller-than-expected practice effects on HVLT-R total and delayed recall, respectively, 12% and 10% of participants displayed smaller-than-expected practice effects on BVMT-R Total and Delayed Recall, respectively, and 16% and 17% of participants displayed smaller-than-expected practice effects on SDMT and TMT-B, respectively). For example, suppose for the HVLT-R Delayed Recall a 70-year-old female participant had a raw score of 5 at observed baseline and a raw score of 6 at observed 1-week, and her predicted 1-week score based on SRB equations was 8.5 (Duff, 2014). This participant has not declined from baseline to follow-up on HVLT-R Delayed Recall, but her degree of improvement over 1 week was smaller than expected based on SRB predictions. As indicated previously, given that the time-frame of repeat test administration was so short and that our sample population is not expected to display acute changes in cognition/treatment response, it is proposed that these findings reflect smaller-than-expected practice effects being observed in our current MCI sample.

Interestingly, while a higher distribution of MCI participants displayed smaller-than-expected practice effects relative to expectations based on the normal distribution of *z* scores, 81% of this sample of MCI participants still displayed the expected degree of practice effect relative to Duff's developmental sample of both cognitively intact and MCI participants. As such, most amnestic MCI participants in this sample benefited from practice to a certain degree. These results support the concept that practice effects, or the capacity to benefit from repeated exposure to information, are impacted by both declarative and procedural memory. While declarative memory is impacted early in the course of amnestic MCI, procedural memory is expected to stay stable in most individuals with amnestic MCI until later in the course of the condition (Duff et al., 2008; Yan & Dick, 2006).

These current findings of smaller-than-expected practice effects in MCI samples are consistent with results of Duff and colleagues (Duff et al., 2018; Duff et al., 2017). Additionally, these findings are consistent with several other studies in the literature reporting an absence or a reduction of practice effects in MCI across a number of cognitive measures and retest intervals (Britt et al., 2011; Calamia et al., 2012; Cooper et al., 2004; Darby et al., 2002; Schrijnemaekers et al., 2006). However, some

ambiguity exists in the literature, as evidenced by other researchers observing improvements on repeated testing in patients with MCI (Duff et al., 2007; Mathews et al., 2014; Yan & Dick, 2006), and these equivocal findings may explain the few measures in our study that failed to show worse outcomes for this MCI sample than predicted based on these SRB prediction equations. For example, BVMT-R Delayed Recall currently failed to show a significantly worse performance compared with prediction, though 10% of participants still displayed smaller-than-expected practice effects. This was surprising given that SRB-based practice effects have been shown to possess significant relationships with amyloid deposition and hippocampal volumes (Duff et al., 2019), brain hypometabolism (Duff et al., 2015), and cognitive decline (Duff et al., 2011); however, other research has failed to show a relationship between SRB-based practice effects for BVMT-R Delayed Recall and hippocampal volumes (Duff et al., 2018). It has been proposed that the relationship between Delayed Recall on the BVMT-R and brain functioning is most noticeable early in the development of cognitive decline (e.g., cognitive change and amyloid deposition), but that the association is diminished later in the course (e.g., atrophy of brain structures; Duff et al., 2018), which may have contributed to our lack of findings in this MCI sample. Despite these variable findings for BVMT-R, these results overall appear to externally validate Duff's (2014) SRB prediction equations for the cognitive measures administered.

Additionally, the results of our secondary analyses partially extend the validation of Duff's (2014) SRB prediction equations. Specifically, we sub-categorized our participants into MCI subtypes based on their baseline performances on non-memory cognitive domains (aMCI-S, aMCI-M for 1–2 non-memory domains, and aMCI-M for 3–4 non-memory domains) to identify whether differential rates of cognitive change or practice effect were observed across amnestic MCI subtypes. Our results indicated that smaller-than-expected practice effects were consistently observed by those amnestic MCI participants with a greater number of impaired cognitive domains for visual immediate and delayed memory (BVMT-R total recall and delayed recall; p values $< .05$). These secondary analyses did not fully extend the validation of Duff's SRB equations because while similar trends were observed for all other measures (p values: .07–.17; see Table 4), they were non-significant.

Because of these equivocal results, we decided to consider "severity of performance" based on test-specific performance as compared to MCI domain impairment. When examining quartiles of test performance for both of the delayed memory measures (HVLT-R Delayed Recall and BVMT-R Delayed Recall) at baseline, we observed that performance severity at baseline is strongly associated with the level of practice effect observed after 1 week for delayed memory measures. For both tasks examined, individuals who performed worse at baseline (i.e., first quartile) had consistently lower z scores relative to individuals from "higher performing" quartiles (e.g., fourth quartile), suggesting that they performed worse relative to SRB prediction equations. When combined with the visual memory secondary analyses, these findings are consistent with research in the literature (Calamia et al., 2012; Cooper et al., 2004) suggesting lower benefits from practice in more severely cognitively compromised samples. For example, Gavett and colleagues (Gavett et al., 2016) found that compared with cognitively intact or MCI samples, patients with Alzheimer's disease displayed disproportionately worse practice effects or improvements on repeat memory testing 1 year apart. This notion has previously been described as "the rich get richer" (Rapport et al., 1997a; Rapport et al., 1997b), such that the stronger baseline performance of the higher performing groups (e.g. aMCI single domain or the fourth quartile performers on the memory measures) appear to have left them poised to benefit from practice effects to a greater extent than the "lower performing" groups (e.g., aMCI-M groups or the first quartile performers), hence the consistently greater z scores observed.

Further, we observed that a greater proportion of participants with aMCI-M (3–4 additional domains impaired) displayed smaller-than-expected practice effects for measures of executive functioning/speeded processing (SDMT and TMT-B; 28–30% of participants) compared to those with aMCI-S (12–13% of participants), and similar non-significant trends were consistently observed for all other measures administered (see Table 5). This discrepant effect for the executive functioning/processing measures in our sample, particularly compared with memory measures, may be explained by multiple factors. First, Suchy and colleagues (Suchy et al., 2011) have demonstrated that practice effects for a measure may be influenced by the test's novelty. As Suchy has tended to use paradigms that incorporate executive capacity as a marker of the novelty effect (Thorgusen et al., 2016), executive functioning measures may be particularly susceptible to practice effects. Additionally, as all participants in our sample met criteria for amnestic MCI, memory dysfunction was common, and therefore it is not necessarily surprising that memory performance, or benefit from practice on memory tasks, did not vary as much as a function of a greater number of non-memory domains being impaired.

The current study is not without limitations. First, these results only inform us about change in participants with amnestic MCI, as we did not have access to clinical samples with either less (cognitively intact) or more severe presentations (e.g., Alzheimer's disease, frontotemporal dementia). Further validation of these prediction equations using a diverse range of neurodegenerative conditions known to affect older adults is warranted to aid in generalizability. Additionally, for the secondary analyses, we possessed a limited number of baseline cognitive measures to categorize our participants into MCI sub-domains. Third, these findings are specific to the cognitive measures administered in this battery over this particular time frame (1 week), and generalization cannot be made to other measures of cognition (e.g., California Verbal Learning Test—II) or different retest intervals (e.g., 1 day, 1 month, or 1 year). Specifically, Calamia and colleagues (Calamia et al., 2012) have previously shown

differential practice effects as a result of measures administered and domains assessed, length of retest interval, use of alternative forms, and diagnostic group. However, as Duff and colleagues (Duff et al., 2010a) previously showed that SRBs may transcend specific tests within a domain, future studies should consider expanding Duff's 2010 work to see how the current results might apply to other tests and intervals. Fourth, these results may not generalize to more heterogeneous participants in regards to premorbid functioning, education, and race. Fifth, regression to the mean's impact on the current results is unclear and worthy of future study. Further, while likely not relevant for MCI samples, the approach taken in the secondary analyses to consider *z* scores for the memory measures as a function of performance ability for each task at baseline may lead to ceiling effects limiting practice effects in participants who performed well at baseline; as a result, readers should take caution applying this approach to healthy control samples.

Finally, an important distinction should be briefly made between statistical and clinical significance in our findings. When examining Table 2, the cognitive battery *z* score values that reached statistical significance were either between −0.48 and − 0.94 or 0.28 and 0.46 (Cohen's *d* values from | 0.53–1.41 |), both of which reflected smaller-than-expected practice effects. While significant at the sample level, when examining the clinical meaning for a particular participant, these *z* score values would be interpreted as displaying an expected level of improvement. Conversely, data from Table 3 emphasizes the individual-level clinical importance of these findings. By displaying high rates of individuals possessing smaller-than-expected practice effects over 1 week (10–29% of the total MCI sample), these latter results lend support to Duff's (2014) SRB prediction equations being capable of identifying clinically meaningful information at the individual level.

Despite these limitations, these results appear to have replicated and extended previous validation of Duff's (2014) SRB prediction equations in MCI samples. Given the potential for practice effects to predict response to intervention (Duff et al., 2010a, b), Alzheimer's-related pathology (Duff et al., 2018; Duff et al., 2014; Galvin et al., 2005; Mormino et al., 2014), and outcomes (Duff et al., 2011; Hassenstab et al., 2015; Machulda et al., 2013) in MCI samples, these current results further support the ability of these SRB prediction equations applied over 1 week to potentially possess diagnostic and prognostic value and inform treatment recommendations. By incorporating the examination of short-term practice effects in clinical settings (e.g., integrated primary care settings), it may be possible develop a more sensitive screen for which individuals are susceptible to negative outcomes associated with neurodegenerative disease, thus reducing the delay between initial screen and intervention.

Funding

The project described was supported by research grants from the National Institutes on Aging: 5R01AG045163. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Aging or the National Institutes of Health.

Conflict of Interest

None declared.

References

Attix, D. K., Story, T. J., Chelune, G. J., Ball, J. D., Stutts, M. L., Hart, R. P., et al. (2009). The prediction of change: Normative neuropsychological trajectories. *The Clinical Neuropsychologist*, *23*(*1*), 21–38. doi: 10.1080/13854040801945078.

Benedict, R. (1997). *Brief visuospatial memory test - Revised: Professional manual*. Lutz, FL: Psychological Assessment Resources, Inc.

Brandt, J., & Benedict, R. (2001). *Hopkins verbal learning test - revised*. Odessa, FL: PAR.

van Britt, W. G.,. 3RD, Hansen, A. M., Bhaskerrao, S., Larsen, J. P., Petersen, F., Dickson, A., et al. (2011). Mild cognitive impairment: Prodromal Alzheimer's disease or something else? *Journal of Alzheimer's Disease*, *27*(*3*), 543–551. doi: 10.3233/JAD-2011-110740.

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(*4*), 543–570.

Chelune, G. (2003). Assessing reliable neuropsychological change. In Franklin, R. (Ed.), *Prediction in forensic and neuropsychology: New approaches to psychometrically sound assessment*. Nahwah, NJ: Erlbaum.

Cooper, D. B., Lacritz, L. H., Weiner, M. F., Rosenberg, R. N., & Cullum, C. M. (2004). Category fluency in mild cognitive impairment: Reduced effect of practice in test-retest conditions. *Alzheimer Disease and Associated Disorders*, *18*(*3*), 120–122.

Crockford, C., Newton, J., Lonergan, K., Madden, C., Mays, I., O'Sullivan, M., et al. (2018). Measuring reliable change in cognition using the Edinburgh cognitive and behavioural ALS screen (ECAS). *Amyotroph Lateral Scler Frontotemporal Degener*, *19*(*1–2*), 65–73. doi: 10.1080/21678421.2017.1407794.

Darby, D., Maruff, P., Collie, A., & McStephen, M. (2002). Mild cognitive impairment can be detected by multiple assessments in a single day. *Neurology*, *59*(*7*), 1042–1046.

Duff, K. (2014). One-week practice effects in older adults: Tools for assessing cognitive change. *The Clinical Neuropsychologist*, *28*(*5*), 714–725. doi: 10.1080/13854046.2014.920923.

Duff, K., Anderson, J. S., Mallik, A. K., Suhrie, K. R., Atkinson, T. J., Dalley, B. C. A., et al. (2018). Short-term repeat cognitive testing and its relationship to hippocampal volumes in older adults. *Journal of Clinical Neuroscience*, *57*, 121–125. doi: 10.1016/j.jocn.2018.08.015.

Duff, K., Atkinson, T. J., Suhrie, K. R., Dalley, B. C., Schaefer, S. Y., & Hammers, D. B. (2017). Short-term practice effects in mild cognitive impairment: Evaluating different methods of change. *Journal of Clinical and Experimental Neuropsychology*, *39*(*4*), 396–407. doi: 10.1080/13803395.2016.1230596.

Duff, K., Beglinger, L. J., Moser, D. J., & Paulsen, J. S. (2010a). Predicting cognitive change within domains. *The Clinical Neuropsychologist*, *24*(*5*), 779–792. doi: 10.1080/13854041003627795.

Duff, K., Beglinger, L. J., Moser, D. J., Schultz, S. K., & Paulsen, J. S. (2010b). Practice effects and outcome of cognitive training: Preliminary evidence from a memory training course. *The American Journal of Geriatric Psychiatry*, *18*(*1*), 91.

Duff, K., Beglinger, L. J., Schultz, S. K., Moser, D. J., McCaffrey, R. J., & Haase, R. F. (2007). Huntington's Study Group, Practice effects in the prediction of long-term cognitive outcome in three patient samples: A novel prognostic index. *Archives of Clinical Neuropsychology*, *22*(*1*), 15–24. doi: 10.1016/j.acn.2006.08.013.

Duff, K., Beglinger, L. J., Van Der, S., Moser, D. J., Arndt, S., Schultz, S. K., et al. (2008). Short-term practice effects in amnestic mild cognitive impairment: Implications for diagnosis and treatment. *International Psychogeriatrics*, *20*(*5*), 986–999. doi: 10.1017/S1041610208007254.

Duff, K., Foster, N. L., & Hoffman, J. M. (2014). Practice effects and amyloid deposition: Preliminary data on a method for enriching samples in clinical trials. *Alzheimer Disease and Associated Disorders*, *28*(*3*), 247–252. doi: 10.1097/WAD.0000000000000021.

Duff, K., Horn, K. P., Foster, N. L., & Hoffman, J. M. (2015). Short-term practice effects and brain hypometabolism: Preliminary data from an FDG PET study. *Archives of Clinical Neuropsychology*, *30*(*3*), 264–270. doi: 10.1093/arclin/acv018.

Duff, K., Lyketsos, C. G., Beglinger, L. J., Chelune, G., Moser, D. J., Arndt, S., et al. (2011). Practice effects predict cognitive outcome in amnestic mild cognitive impairment. *The American Journal of Geriatric Psychiatry*, *19*(*11*), 932–939. doi: 10.1097/JGP.0b013e318209dd3a.

Duff, K., Schoenberg, M. R., Patton, D., Mold, J., Scott, J. G., & Adams, R. L. (2004). Predicting change with the RBANS in a community dwelling elderly sample. *Journal of the International Neuropsychological Society*, *10*(*6*), 828–834.

Duff, K., Schoenberg, M. R., Patton, D., Paulsen, J. S., Bayless, J. D., Mold, J., et al. (2005). Regression-based formulas for predicting change in RBANS subtests with older adults. *Archives of Clinical Neuropsychology*, *20*(*3*), 281–290. doi: 10.1016/j.acn.2004.07.007.

Duff, K., Suhrie, K. R., Dalley, B. C. A., Anderson, J. S., & Hoffman, J. M. (2019). External validation of change formulae in neuropsychology with neuroimaging biomarkers: A methodological recommendation and preliminary clinical data. *The Clinical Neuropsychologist*, *33*(*3*), 478–489. doi: 10.1080/13854046.2018.1484518.

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics*, *19*(*4*), 27–29.

Galvin, J. E., Powlishta, K. K., Wilkins, K., McKeel, D. W., Jr., Xiong, C., Grant, E., et al. (2005). Predictors of preclinical Alzheimer disease and dementia: A clinicopathologic study. *Archives of Neurology*, *62*(*5*), 758–765. doi: 10.1001/archneur.62.5.758.

Gavett, B. E., Ashendorf, L., & Gurnani, A. S. (2015). Reliable change on neuropsychological tests in the uniform data set. *Journal of the International Neuropsychological Society*, *21*(*7*), 558–567. doi: 10.1017/S1355617715000582.

Gavett, B. E., Gurnani, A. S., Saurman, J. L., Chapman, K. R., Steinberg, E. G., Martin, B., et al. (2016). Practice effects on story memory and list learning tests in the neuropsychological assessment of older adults. *PLoS One*, *11*(*10*), e0164492. doi: 10.1371/journal.pone.0164492.

Hammers, D., Duff, K., & Chelune, G. (2015). Assessing change of cognitive trajectories over time in later life. In Pachana, N. A., & Laidlaw, K. (Eds.), *Oxford handbook of clinical geropsychology*. Oxford, England: Oxford University Press.

Hassenstab, J., Ruvolo, D., Jasielec, M., Xiong, C., Grant, E., & Morris, J. C. (2015). Absence of practice effects in preclinical Alzheimer's disease. *Neuropsychology*, *29*(*6*), 940–948. doi: 10.1037/neu0000208.

Hinton-Bayre, A. D. (2010). Deriving reliable change statistics from test-retest normative data: Comparison of models and mathematical expressions. *Archives of Clinical Neuropsychology*, *25*(*3*), 244–256. doi: 10.1093/arclin/acq008.

Lezak, M., Howieson, D., Bigler, E., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). New York: Oxford University Press.

Machulda, M. M., Pankratz, V. S., Christianson, T. J., Ivnik, R. J., Mielke, M. M., Roberts, R. O., et al. (2013). Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo clinic study of aging. *The Clinical Neuropsychologist*, *27*(*8*), 1247–1264. doi: 10.1080/13854046.2013.836567.

Mathews, M., Abner, E., Kryscio, R., Jicha, G., Cooper, G., Smith, C., et al. (2014). Diagnostic accuracy and practice effects in the national Alzheimer's coordinating center uniform data set neuropsychological battery. *Alzheimers Dement*, *10*(*6*), 675–683. doi: 10.1016/j.jalz.2013.11.007.

McSweeny, A., Naugle, R. I., Chelune, G. J., & Luders, H. (1993). "T-scores for change:" An illustration of a regression approach to depicting change in clinical neuropsychology. *The Clinical Neuropsychologist*, *7*, 300–312.

Mormino, E. C., Betensky, R. A., Hedden, T., Schultz, A. P., Amariglio, R. E., Rentz, D. M., et al. (2014). Synergistic effect of beta-amyloid and neurodegeneration on cognitive decline in clinically normal individuals. *JAMA Neurology*, *71*(*11*), 1379–1385. doi: 10.1001/jamaneurol.2014.2031.

Randolph, C. (2012). *Repeatable battery for the assessment of neuropsychological status*. Bloomington, MN: The Psychological Corporation.

Rapport, L. J., Axelrod, B. N., Theisen, M. E., Brines, D. B., Kalechstein, A. D., & Ricker, J. H. (1997a). Relationship of IQ to verbal learning and memory: Test and retest. *Journal of Clinical and Experimental Neuropsychology*, *19*(*5*), 655–666. doi: 10.1080/01688639708403751.

Rapport, L. J., Brines, D., Axelrod, B., & Theisen, M. E. (1997b). Full scale IQ as mediator of practice effects: The rich get richer. *The Clinical Neuropsychologist*, *11*(*4*), 375–380.

Reitan, R. (1992). *Trail making test: Manual for administration and scoring*. Tucson, AZ: Reitan Neuropsychology Laboratory.

Rinehardt, E., Duff, K., Schoenberg, M., Mattingly, M., Bharucha, K., & Scott, J. (2010). Cognitive change on the repeatable battery of neuropsychological status (RBANS) in Parkinson's disease with and without bilateral subthalamic nucleus deep brain stimulation surgery. *The Clinical Neuropsychologist*, *24*(*8*), 1339–1354. doi: 10.1080/13854046.2010.521770.

Sanchez-Benavides, G., Pena-Casanova, J., Casals-Coll, M., Gramunt, N., Manero, R. M., Puig-Pijoan, A., et al. (2016). One-year reference norms of cognitive change in Spanish old adults: Data from the NEURONORMA sample. *Archives of Clinical Neuropsychology*, *31*(*4*), 378–388. doi: 10.1093/arclin/acw018.

Schrijnemaekers, A. M., de Jager, C. A., Hogervorst, E., & Budge, M. M. (2006). Cases with mild cognitive impairment and Alzheimer's disease fail to benefit from repeated exposure to episodic memory tests as compared with controls. *Journal of Clinical and Experimental Neuropsychology*, *28*(*3*), 438–455. doi: 10.1080/13803390590935462.

Smith, A. (1973). *Symbol digit modalities test*. Los Angeles, CA: Western Psychological Services.

Stein, J., Luppa, M., Brahler, E., Konig, H. H., & Riedel-Heller, S. G. (2010). The assessment of changes in cognitive functioning: Reliable change indices for neuropsychological instruments in the elderly - a systematic review. *Dementia and Geriatric Cognitive Disorders*, *29*(*3*), 275–286. doi: 10.1159/000289779.

Suchy, Y., Kraybill, M. L., & Franchow, E. (2011). Practice effect and beyond: Reaction to novelty as an independent predictor of cognitive decline among older adults. *Journal of the International Neuropsychological Society*, *17*(*1*), 101–111. doi: 10.1017/S135561771000130X.

Thorgusen, S. R., Suchy, Y., Chelune, G. J., & Baucom, B. R. (2016). Neuropsychological practice effects in the context of cognitive decline: Contributions from learning and task novelty. *Journal of the International Neuropsychological Society*, *22*(*4*), 453–466. doi: 10.1017/S1355617715001332.

Wilkinson, G. S., & Robertson, G. J. (2006). *WRAT 4: Wide range achievement test, professional manual*. Lutz, FL: Psychological Assessment Resources, Inc.

Yan, J. H., & Dick, M. B. (2006). Practice effects on motor control in healthy seniors and patients with mild cognitive impairment and Alzheimer's disease. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, *13*(*3–4*), 385–410. doi: 10.1080/138255890969609.